Chinese Physical Society | IOP Publishing

CPB
Chinese Physics B

http://cpb.iphy.ac.cn https://iopscience.iop.org/journal/1674-1056

DSAS: A new macromolecular substructure solution program based on the modified phase-retrieval algorithm

Xingke Fu(付兴科), Zhenxi Tan(谭振希), Zhi Geng(耿直), Qian Liu(刘茜), and Wei Ding(丁玮)

Citation: Chin. Phys. B, 2024, 33 (**5**): 056102. DOI: 10.1088/1674-1056/ad3c33

Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Surface structure modification of $ReSe_2$ nanosheets via carbon ion irradiation

Mei Qiao(乔梅), Tie-Jun Wang(王铁军), Yong Liu(刘泳), Tao Liu(刘涛), Shan Liu(刘珊), and Shi-Cai Xu(许士才)

Chin. Phys. B, 2023, 32 (**2**): 026101. DOI: 10.1088/1674-1056/ac7297

Structure, phase evolution and properties of Ta films deposited using hybrid high-power pulsed and DC magnetron co-sputtering

Min Huang(黄敏), Yan-Song Liu(刘艳松), Zhi-Bing He(何智兵), and Yong Yi(易勇)

Chin. Phys. B, 2022, 31 (**6**): 066101. DOI: 10.1088/1674-1056/ac43a9

Ab initio study on crystal structure and phase stability of ZrC_2 under high pressure

Yong-Liang Guo(郭永亮), Jun-Hong Wei(韦俊红), Xiao Liu(刘潇), Xue-Zhi Ke(柯学志), and Zhao-Yong Jiao(焦照勇)

Chin. Phys. B, 2021, 30 (**1**): 016101. DOI: 10.1088/1674-1056/abb3e7

Anti-oxidation characteristics of Cr-coating on surface of Ti-45Al-8.5Nb alloy by plasma surface metallurgy technique

Bing Zhou(周兵), Ya-Rong Wang(王亚榕), Ke Zheng(郑可), Yong Ma(马永), Yong-Sheng Wang(王永胜), Sheng-Wang Yu(于盛旺), and Yu-Cheng Wu(吴玉程)

Chin. Phys. B, 2020, 29 (**12**): 126101. DOI: 10.1088/1674-1056/aba9c2

Development of “Parameter space screening”-based single-wavelength anomalous diffraction phasing and structure determination pipeline

Wei Ding(丁玮), Xiao-Ting Wang(王小婷), Yang-Yang Yi(易阳旸)

Chin. Phys. B, 2019, 28 (**11**): 116101. DOI: 10.1088/1674-1056/ab43bd

----

COMPUTATIONAL PROGRAMS FOR PHYSICS

DSAS: A new macromolecular substructure solution program based on the modified phase-retrieval algorithm

Xingke Fu(付兴科)[1,4,†], Zhenxi Tan(谭振希)[2,†], Zhi Geng(耿直)[3,4,‡],
Qian Liu(刘茜)[2,§], and Wei Ding(丁玮)[1,4,5,¶]

[1] *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*
[2] *Beijing Yunlu Technology Co., Ltd., Beijing* 100161, *China*
[3] *Beijing Synchrotron Radiation Facility, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing* 100049, *China*
[4] *School of Physical Sciences, University of Chinese Academy of Sciences, Beijing* 100049, *China*
[5] *Songshan Lake Materials Laboratory, Dongguan* 523808, *China*

(Received 15 March 2024; revised manuscript received 8 April 2024; accepted manuscript online 9 April 2024)

Considering the pivotal role of single-wavelength anomalous diffraction (SAD) in macromolecular crystallography, our objective was to introduce DSAS, a novel program designed for efficient anomalous scattering substructure determination. DSAS stands out with its core components: a modified phase-retrieval algorithm and automated parameter tuning. The software boasts an intuitive graphical user interface (GUI), facilitating seamless input of essential data and real-time monitoring. Extensive testing on DSAS has involved diverse datasets, encompassing proteins, nucleic acids, and various anomalous scatters such as sulfur (S), selenium (Se), metals, and halogens. The results confirm DSAS's exceptional performance in accurately determining heavy atom positions, making it a highly effective tool in the field.

Keywords: DSAS, single-wavelength anomalous diffraction, automated parameters settings, phase-retrieval algorithm, substructure determination

PACS: 61.05.–a **DOI:** 10.1088/1674-1056/ad3c33

1. Introduction

It is a fundamental axiom of biology that the three-dimensional structure of a protein determines its function. Understanding function through structure is a primary goal of structural biology.[1] Exhilaratingly, x-ray crystallography can resolve the structures of biological macromolecules (such as proteins) at atomic resolution, greatly enriching the Protein Data Bank (PDB).[2] Single-wavelength anomalous diffraction (SAD),[3] being one of the most important crystallographic structural determination methods, continues to play a significant role in exploring unknown macromolecular structures without homologues. Especially, SAD remains the dominant method for determining the structures of nucleic acids[4,5] and protein.[6,7] Based on the type of anomalous scatters, there are currently several commonly used variants of SAD: S-SAD, Se-SAD (labeling proteins with selenomethionine), and X-SAD (iodine, bromine, or metal ions).

The SAD method can be divided into four main steps: determining the anomalous substructure, generating protein phases, density modification, and model building. The identification of anomalous substructure is the first and most crucial step in SAD phasing and it is still the main bottleneck in the SAD phasing process.[8]

A traditional method is the tangent formula-based direct methods, which directly estimate phases from the relationship betweeen the intensities and phases. Of note, direct methods are typically incorporated into a dual-space iteration framework,[9] in which the direct method is used for phase refinement in reciprocal space, and the prior knowledge of biological or chemical information (such as atomicity) is used for density constrain in real space. A typical and authoritative direct method-based program is SHELXD,[10] which uses the Patterson-based seeding instead of random phases or positions for the initial phase estimation. Since the initial heavy atom coordinate positions are consistent with the Patterson superposition minimum function (PSMF),[11] the efficiency of the dual-space iterative method can be improved by roughly an order of magnitude. In 2015, Bunkóczi *et al.* recommended a SAD likelihood scoring function to rank candidate substructures derived from the anomalous difference Patterson function.[12] This approach enabled them to identify missing sites and ultimately achieve a nearly complete solution.

Moreover, several algorithms and programs based on novel computational frameworks have also been explored for the determination of heavy-atom substructures. In 2019, Hu *et al.* introduced noise and artifact suppression using resampling

[†] These authors contributed equally to the work.
[‡] Corresponding author. E-mail: gengz@ihep.ac.cn
[§] Corresponding author. E-mail: ss818cc@hotmail.com
[¶] Corresponding author. E-mail: dingwei@iphy.ac.cn

(NASR) method from nuclear magnetic resonance into macromolecular crystallography.[13] This method can improve the signal-to-noise ratio (SNR) of the difference Patterson map, which facilitates real space-based substructure determination, such as RSPS program.[14] In 2022, Rius and Torrelles developed a SAD-SAMR algorithm[15] that has been implemented in a modified version of XLENS_v1.[16] This algorithm is a density-based phasing algorithm that incorporates a peakness-enhancing *ipp* (inner-pixel preservation) procedure, which features two Fourier iterations into one phase refinement cycle.

What is more, it is also worth noting that the phase-retrieval algorithms, which are one of the most commonly used structure determination algorithms in chemical crystallography, have been also extended to macromolecular substructure determination. For example, in 2008, Dumas and van der Lee successfully solved a dataset containing 120 heavy atoms using the charge flipping (CF) based program SUPERFLIP, observing the presence of sixfold noncrystallographic symmetry between these sites.[17] Furthermore, in 2018, Skubák published a new program PRASA for substructure determination based on a new adaptation of the phase-retrieval algorithm, where the relaxed alternating averaged reflection (RAAR) phase-retrieval algorithm was firstly used to solve macromolecular substructure, and exhibited superior performance compared to CF algorithm.[18] And the phase-retrieval algorithm is also working in dual-space framework, but it just involves the magnitude constraints in reciprocal space and low-density perturbations in real space.

Recently, we introduced a modified phase-retrieval algorithm for tackling anomalous scattering substructures in macromolecular crystals. In contrast to the conventional RAAR algorithm, our modification integrates the π-half phase perturbation in the CF algorithm for weak reflections and seamlessly incorporates a direct method-based tangent formula into dual-space iterations. This enhancement not only renders the algorithm more adaptable but also significantly boosts its robustness, enabling it to effectively resolve challenging heavy-atom substructures even under unfavorable conditions.

We have developed a user-friendly program called DSAS (Dual-Space Algorithm for Anomalous Substructures) based on the modified phase-retrieval algorithm. It automates all steps from processing anomalous scattering experimental data to the final solution. We have successfully applied DSAS to a diverse range of macromolecular data sets, encompassing various proteins, nucleic acids, and different heavy atom types. This demonstrates the powerful versatility and ease of use of DSAS, making it accessible to a wide range of users.

2. Method

2.1. Introduction of the modified phase-retrieval algorithm

The DSAS is designed based on the modified phase-retrieval algorithm, which harmoniously combines the π-half phase perturbation for weak reflections and the direct method-based tangent formula for strong reflections within the RAAR algorithm framework.[19,20] The phase-retrieval algorithm belongs to perturbation-based dual-space iterative algorithm. It achieves convergence by balancing constraints and perturbations in real and reciprocal spaces. The calculated electron density in cycle n can expressed as

$$\rho_n = \Theta_{\mathrm{D}} \mathcal{F} \Theta_{\mathrm{M}} \mathcal{F}^{-1} \rho_{n-1}, \tag{1}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote forward and inverse Fourier transforms, Θ_{M} and Θ_{D} correspond to the constraint operators in reciprocal and real spaces, respectively. One refinement iteration can be described as follows.

The initial electron density map can be either the modified density map from the previous cycle or the 0th cycle density map which is constructed from random initial phases and anomalous or dispersive difference amplitudes from SAD experiment. In reciprocal space, the constraints can be divided into amplitude constraints and phase constraints. For magnitude constraints, the calculated amplitudes $\boldsymbol{F}_h^{\mathrm{c}}$ are replaced by experimental amplitudes $|\boldsymbol{F}_h^{\mathrm{o}}|$, while the phases remain unchanged, and unobserved amplitudes remain unchanged, as shown below:

$$\Theta_{\mathrm{M}}\left(\boldsymbol{F}_h^{\mathrm{c}}\right) = \begin{cases} \dfrac{|\boldsymbol{F}_h^{\mathrm{o}}|}{|\boldsymbol{F}_h^{\mathrm{c}}|}\boldsymbol{F}_h^{\mathrm{c}}, & \text{if } \boldsymbol{h} \in H_{\mathrm{obs}}, \\ \boldsymbol{F}_h^{\mathrm{c}}, & \text{otherwise}, \end{cases} \tag{2}$$

where H_{obs} is the set of reflections $\boldsymbol{h}$ for which the experimental amplitudes are known. And the phase constraints involve the π-half phase perturbation for weak reflections[21] (see Eq. (3)) and enforcing the direct-method-based tangent formula[22] for strong reflections. The phases of the reflections $\boldsymbol{h}$ considered to be weak are offset by a constant $\Delta\varphi = \pi/2$, which is so-called π-half variant,

$$\varphi_h = \begin{cases} \varphi_i^h + \dfrac{\pi}{2}, & \text{if } \boldsymbol{h} \in H_{\mathrm{weak}}, \\ \varphi_i^h, & \text{otherwise}, \end{cases} \tag{3}$$

where φ_i^h denotes the calculated phases at current iteration and H_{weak} is the set of weak reflections where w_{best} percent (20%–50%) of the experimental amplitudes will be considered weak based on their experimental amplitudes. Besides, a specified number (N_{TF}) of strongest reflections will be further refined by the direct-method tangent formula.

In real space, we made a slight modification to the real-space constraints of the RAAR algorithm, which is used as the real-space density modification of the modified phase-retrieval algorithm. Here, only the densities with small values in the density map are modified, while the previously calculated density map is also considered. In addition, a positivity constraint is introduced in real space, which takes the absolute values of the densities before and after the real-space density modification. The refinement $\varTheta_{\mathrm{D}}$ can be express as

$$\begin{aligned}
&\rho_i' = |\mathcal{F}\varTheta_{\mathrm{M}}\mathcal{F}^{-1}\rho_i^{n-1}|,\\
&\rho_i^n = \begin{cases} \rho_i' & \text{for } 2\rho_i - \rho_i^{n-1} > \delta,\\ \beta\rho_i^{n-1} + (1-2\beta)\rho_i' & \text{for } 2\rho_i - \rho_i^{n-1} < \delta, \end{cases}\\
&\rho_i^n = |\rho_i^n|,
\end{aligned} \tag{4}$$

where β is a coefficient of relaxation term, δ signifies the threshold of electron density values, ρ_i^{n-1} denotes the calculated density map at the last iteration.

2.2. Workflow of DSAS

The flow-chart of the DSAS is shown in Fig. 1 and the details are described below.

Step 1 Experimental data preprocessing

Firstly, the input experimental data with anomalous scattering signals (including F^+/F^- or I^+/I^-) would be converted to heavy-atom substructure factors F_{A}, $|F_{\mathrm{A}}| = |F^+| + |F^-|$ using SHELXC.[23] Then, F_{A} are further normalized to pseudo-normalized amplitudes E_{A} using ECALC from CCP4 suite[24] for the tangent formula and the phase-retrieval algorithm.[25]

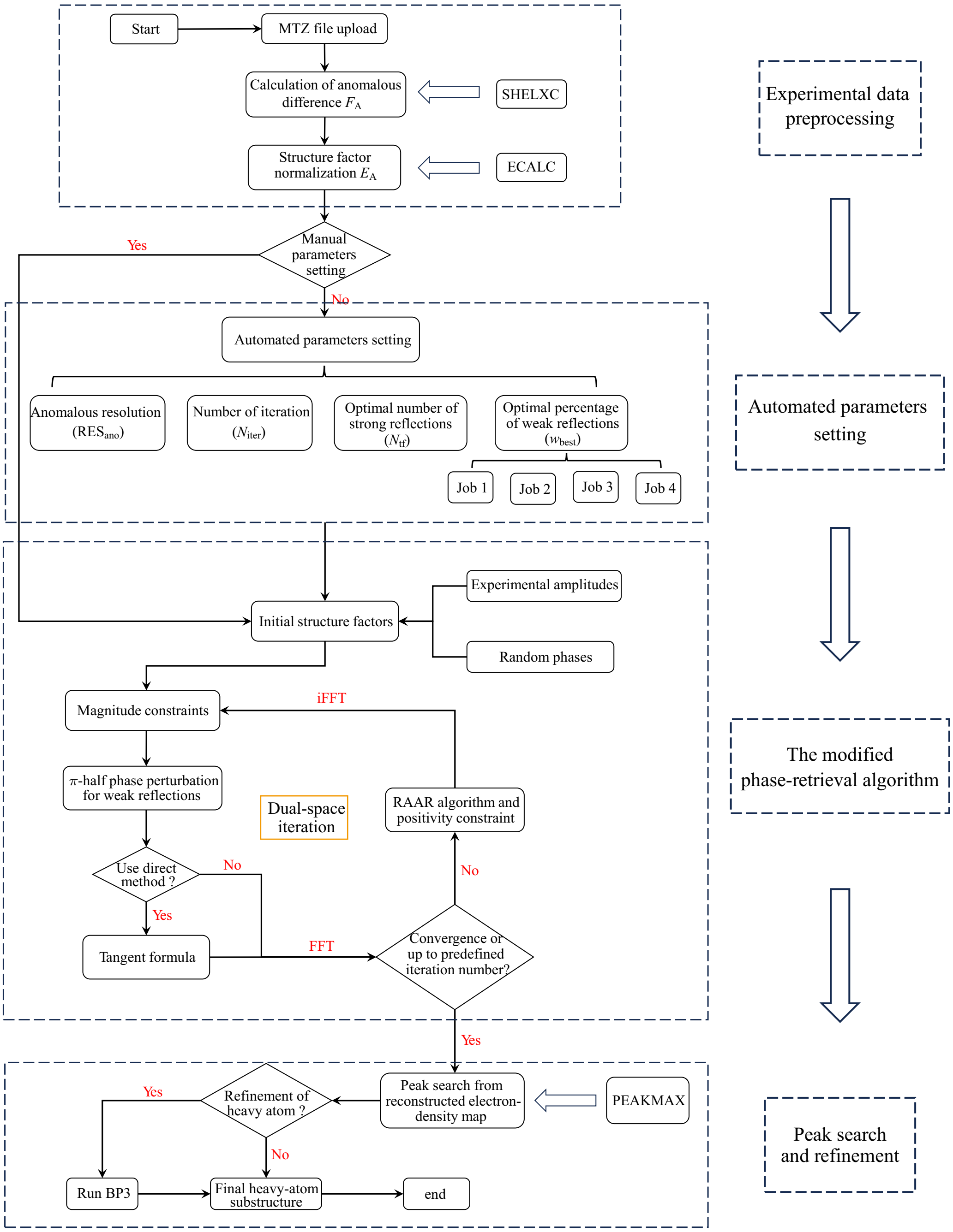

Start
MTZ file upload
Calculation of anomalous difference F_{A}
SHELXC
Structure factor normalization E_{A}
ECALC
Experimental data preprocessing
Yes
Manual parameters setting
No
Automated parameters setting
Anomalous resolution (RES$_{\mathrm{ano}}$)
Number of iteration (N_{iter})
Optimal number of strong reflections (N_{tf})
Optimal percentage of weak reflections (w_{best})
Job 1
Job 2
Job 3
Job 4
Automated parameters setting
Experimental amplitudes
Initial structure factors
Random phases
Magnitude constraints
iFFT
π-half phase perturbation for weak reflections
Dual-space iteration
RAAR algorithm and positivity constraint
Use direct method ?
No
Yes
No
Tangent formula
FFT
Convergence or up to predefined iteration number?
The modified phase-retrieval algorithm
Yes
Yes
Refinement of heavy atom ?
Peak search from reconstructed electron-density map
PEAKMAX
No
Run BP3
Final heavy-atom substructure
end
Peak search and refinement

Fig. 1. The flow-chart of the DSAS.

Step 2 Automated parameters setting

We have identified several important parameters for DSAS, including the relaxation parameter β, the electron-density threshold δ, anomalous resolution $\mathrm{RES}_{\mathrm{ano}}$, the number of iterations for each trial N_{iter}, the number of strong reflections N_{TF} and the optimal percentage of weak reflections w_{best}. To streamline the process, we have developed an automated parameter-setting method that self-adaptively calculates the appropriate parameters for each SAD data set based on the anomalous scattering information.

In our test, we found a constant value of 0.82 for β and the dynamical threshold charge δ in every cycle to stabilize 13% of the density pixels which work well in most cases. The anomalous resolution $\mathrm{RES}_{\mathrm{ano}}$ refers to the truncated resolution of anomalous difference data, which will be used to remove some invalid reflections with weak anomalous signals at high resolution, improving the signal-to-noise ratio of the data. Here, the ratio of the anomalous difference to its standard deviation ($|\Delta F|/\sigma(\Delta F) = 1.2$) is used to estimate the $\mathrm{RES}_{\mathrm{ano}}$.[26] The N_{iter} is set to 500 (or 750 for $\mathrm{RES}_{\mathrm{ano}} < 3.8$ Å), and we hope to compensate for the coordinate errors caused by low resolution by increasing the number of iterations.

The number of strong reflections N_{TF} is contingent upon the number of observed reflections. N_{TF} is set to 500 when the observed reflection count is below 200, to 100 when the count is above 2000 but below 500, to 1300 when the count is above 5000 but below 8000, and to 1500 in all other cases. The w_{best} is determined through the parameter space screening method. Four jobs are run, each with 10 trials, to test different values of w_{best} (0.2, 0.3, 0.4, and 0.5). The optimal value is determined by considering their convergence trends.

Step 3 Implementation of the modified phase-retrieval algorithm

Once the parameters and experimental data required for the modified phase-retrieval algorithm are prepared, we begin to implement this substructure solution with 400 trials for each SAD data set. The potential solutions are identified by the Pearson correlation coefficient between E_{o} and E_{c},

$$\mathrm{CC} = \frac{n\sum E_{\mathrm{o}}E_{\mathrm{c}} - \sum E_{\mathrm{o}}\sum E_{\mathrm{c}}}{\sqrt{\left[n\sum E_{\mathrm{o}}^2 - \left(\sum E_{\mathrm{o}}\right)^2\right]\left[n\sum E_{\mathrm{c}}^2 - \left(\sum E_{\mathrm{c}}\right)^2\right]}}, \tag{5}$$

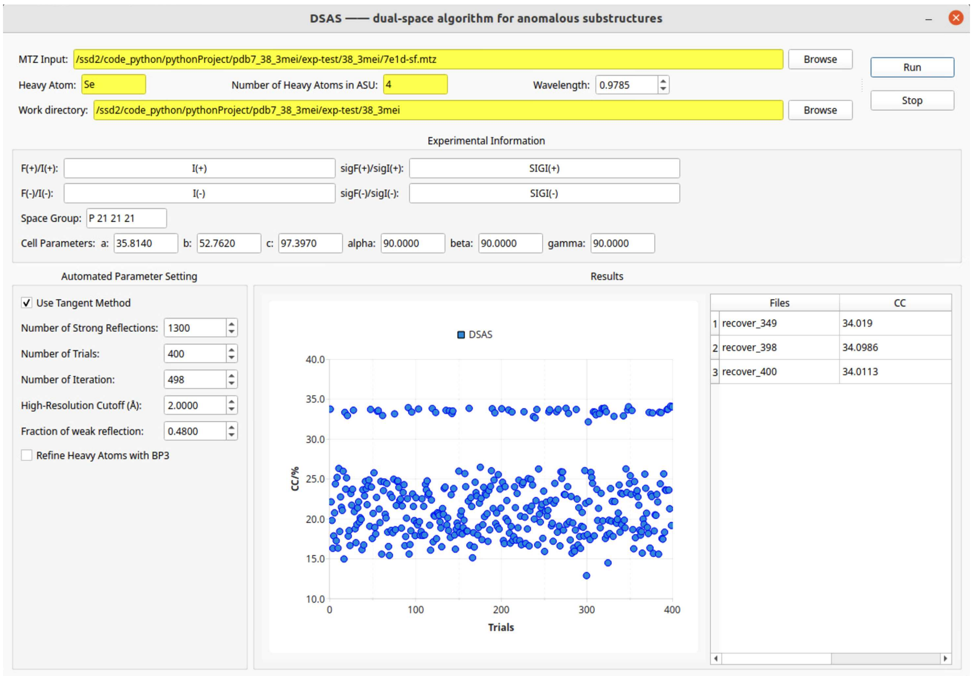
where E_{o} and E_{c} represent the observed and calculated normalized amplitudes, and n represents the number of observed reflections.

Step 4 Peak search and refinement

The final solution of substructure can be obtained from the asymmetric unit of the Fourier map with the best CC using PEAKMAX program from CCP4 suite. An optional step is to use BP3[27] to refine the parameters of the potential substructure, such as 3D atomic coordinates, occupancy, and temperature factor. The default number of heavy atoms output is two greater than the number of deposited heavy atoms. Finally, the files containing heavy atom information are output in PDB file format.

2.3. Graphical user interface

The DSAS has a user-friendly GUI (as shown in Fig. 2) written in PyQt5. It mainly consists of four panels: inputting panel, experimental information panel, automated parameters setting panel, and results panel.



DSAS —— dual-space algorithm for anomalous substructures

MTZ Input: /ssd2/code_python/pythonProject/pdb7_38_3mei/exp-test/38_3mei/7e1d-sf.mtz

Heavy Atom: Se

Number of Heavy Atoms in ASU: 4

Wavelength: 0.9785

Work directory: /ssd2/code_python/pythonProject/pdb7_38_3mei/exp-test/38_3mei

Experimental Information

F(+)/I(+): I(+)

sigF(+)/sigI(+): SIGI(+)

F(-)/I(-): I(-)

sigF(-)/sigI(-): SIGI(-)

Space Group: P 21 21 21

Cell Parameters: a: 35.8140 b: 52.7620 c: 97.3970 alpha: 90.0000 beta: 90.0000 gamma: 90.0000

Automated Parameter Setting

☑ Use Tangent Method

Number of Strong Reflections: 1300

Number of Trials: 400

Number of Iteration: 498

High-Resolution Cutoff (Å): 2.0000

Fraction of weak reflection: 0.4800

□ Refine Heavy Atoms with BP3

Results

	Files	CC
1	recover_349	34.019
2	recover_398	34.0986
3	recover_400	34.0113

Fig. 2. GUI of the DSAS. It contains four panels: inputting panel, experimental information panel, automated parameters setting panel, and results panel.

The inputting panel is used to upload MTZ file. Besides, the type of heavy atom and the number of heavy atoms in the asymmetric unit (ASU) can also be input. The number of heavy atoms does not need to be very accurate, as it will only be used for the final peak search and will not affect the calculation results. It is recommended to set it slightly larger if it is unknown. The inputting "wavelength" is the x-ray wavelength of SAD experiment. The inputting "work directory" is the path to the output files of the DSAS. The experimental information panel exhibits the anomalous information, space group and cell parameters. The automated parameters setting panel outputs the parameters automatically determined using the automated parameters setting method. By default, the tangent formula is used to refine the phases every 20 cycles after 100 Fourier iterations, and BP3 is not used to correct the final heavy-atom parameters. Of course, these parameters can also be set manually. The results panel displays the CC values of each trial in real time and finally outputs the three results with the highest CC values.

3. Results and discussion

Many previously unknown crystal structures have been automatically solved by DSAS using default protocols. However, this study presents only 14 representative examples in Table 1, comprising 5 SeMet-SAD data, 3 halogen-SAD data, 3 sulfur-SAD data, and 3 RNA/DNA structures. The quality of the output results is evaluated using three indicators: the number of sites found in the asymmetric unit (a.u.), the mean error of calculated heavy atoms, and the root mean square deviation of the calculated heavy-atom substructure. These indicators values are generated by the SITCOM program,[28] which compares the substructure atoms with the actual heavy atoms extracted from the reference PDB coordinates. An overview of the results and the effective parameters of the test cases is presented in Table 2.

Overall, in all test cases, over 70% of heavy atom sites were successfully placed using the default protocol. Particularly in the SeMet-SAD, halogen-SAD, and RNA/DNA test cases, the correct proportion was close to 100%. Even when dealing with the notoriously challenging native sulfur-SAD, DSAS demonstrated an impressive accuracy of 72%–83%. What is more, two popular site search programs SHELXD and *Phenix.hyss* were also employed for substructure determination using default parameters, with the results presented in Table 2. As observed, in most case, DSAS demonstrated superior statistical results. This highlights the comparable efficiency of DSAS to state-of-the-art substructure determination tools.

Table 1. Diffraction data used in the case studies.

Test case	PDB entry	Type	Heavy atom	[1]a.s.	[2]$d_{\min}$ (Å)	[3]No. in a.u.	[4]λ (Å)	Space group	Unit cell	
									a, b, c (Å)	α, β, γ (°)
Test case 1	5COM	protein	Se	40.05	1.85	4	0.980	$C\,1\,2\,1$	127.05, 50.3, 91.72	90, 118.7, 90
	5XKY	protein	Se	15.04	2.303	9	0.978	$P\,2_12_12_1$	61.221, 72.375, 88.341	90, 90, 90
	8EZS	protein	Se	11.06	2.473	8	0.978	$P\,4_12_12$	44.854, 44.854, 392.053	90, 90, 90
	6HP5	protein	Se	13.54	2.28	30	0.979	$P\,4_1\,2_1\,2$	94.34, 94.34, 249.95	90, 90, 90
	8PX1	protein	Se	6.93	2.1	20	2.755	$P\,1\,2_11$	96.08, 75.02, 101.89	90, 109.52, 90
Test case 2	4RYM	protein	I	9.95	2.8	3	2.074	$P\,2_12_12_1$	33.356, 49.538, 99.213	90, 90, 90
	2RIW	protein	I	33.63	2.038	4	1.542	$P\,2_12_12$	173.594, 42.164, 55.98	90, 90, 90
	5IO8	protein	I	10.35	2.192	14	1.771	$P\,2_12_12$	60.115, 60.277, 62.603	90, 90, 90
Test case 3	8PWN	protein	S	5.51	2.4	21	2.755	$C\,2\,2\,2_1$	39.31, 179.64, 139.54	90, 90, 90
	7O51	protein	S	6.98	2.2	18	2.059	$P\,4_12_12$	58.52, 58.52, 151.3	90, 90, 90
	5II7	protein	S	9.79	1.66	18	1.542	$P\,2_12_12_1$	45.196, 51.195, 80.417	90, 90, 90
Test case 4	*5LQO	RNA	Br	10.42	1.87	2	0.9196	$P\,4_32_12$	33.469, 33.469, 113.766	90, 90, 90
	*3MEI	RNA	Br	28.56	1.968	2	0.9197	$P\,1\,2_11$	32.928, 35.564, 41.985	90, 100.1, 90
	*7OW0	DNA + RNA	Zn	20.36	1.548	3	1.283	$P\,4_32_12$	31.85, 31.85, 91.66	90, 90, 90

[1] a.s.: anomalous signal, the mean peak height of a normalized anomalous difference Fourier map.

[2] $d_{\min}$: the experimental diffraction limit.

[3] No. in a.u.: the number of sites in the asymmetric unit (a.u.) from published macromolecular structure.

[4] λ: the wavelength of x-ray.

*: the cases that cannot be solved by the MR-model-building method.

3.1. Contribution of phase constraints in the modified phase-retrieval algorithm

As shown in Fig. 3, with the introduction of the π-half phase perturbation and direct methods, the efficiency of the modified phase-retrieval algorithm is significantly higher compared to the standard RAAR algorithm. We also noticed that the use of the direct method will reduce the CC value (as shown by the red dots in Fig. 3). One possible reason is that the introduction of the tangent formula will help to escape the stagnate at false local minima and accelerate the convergence. However, continuous use is equivalent to introducing a large perturbation to this dual-space algorithm, which is prone to divergence.

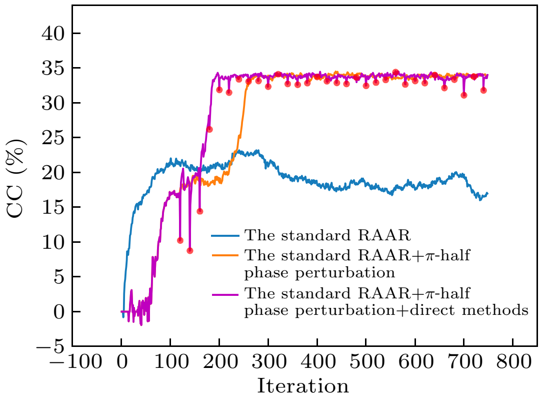


Fig. 3. Comparison of the RAAR algorithms with and without phase constraints (the π-half variant and tangent formula) across 750 Fourier iterations for protein with PDB entry 7E1D, all starting with the same random phase values. The distinct curves with different colors illustrate the variations in CC values during the iterative process of different algorithm strategies. The red dots represent the use of tangent formula.

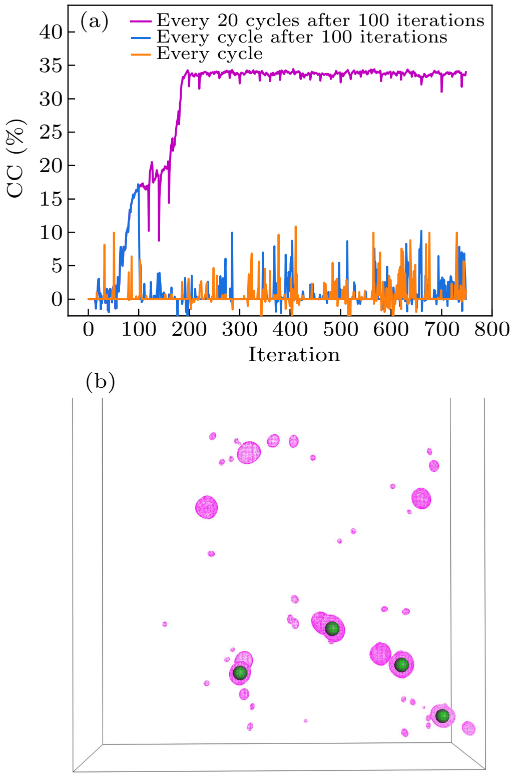


Fig. 4. Comparison of strategy of direct-method refinement. (a) With different direct-method implementation strategies, the CC value as a function of the iterative correction cycles. The strategy of every 20 cycles after 100 iterations is indicated in purple; the strategy of every cycle after 100 iterations is indicated in blue; the strategy of every cycle is indicated in orange. (b) The converged electron density map from the strategy of every 20 cycles after 100 iterations. The green balls represent 4 Se atoms in the asymmetric unit from the PDB-deposited structure. The electron density maps are contoured at 4σ.

Of particular note, unlike Coelho's strategy[22] of constraining the phases during the whole iteration process, here, the phases are refined using the tangent formula every 20 cycles after 100 Fourier iterations. As shown in Fig. 4(a), the strategy of direct-method refinement every 20 cycles after 100 Fourier iterations results in a faster convergence compared to the other two strategies of continuously using the direct-method refinement. And the converged electron density map has a high signal-to-noise ratio and is consistent with the coordinate positions of the reference substructure (Fig. 4(b)).

3.2. Substructure determination for SeMet-SAD and Halogen-SAD data

Test cases 1 and 2 are representative of typical SeMet-SAD and Halogen-SAD data sets, encompassing a range of resolutions, space groups, and site numbers (Table 1). In the SeMet-SAD method, SeMet is incorporated into the protein during its expression, whereas the halogen-SAD method employs heavy halogen atoms, typically through derivatization of native amino acid side chains with halogen-containing compounds. Both SeMet and halogen possess large atomic numbers, and when the x-ray wavelength approximates the K-absorption edge, robust anomalous scattering signals are always strong enough for phasing the diffraction data. Table 2 showcases the exceptional performance of DSAS in the SeMet-SAD and Halogen-SAD data sets, accurately identifying nearly complete substructures with minimal positional deviation from the reference structure. However, it should be noted that for 8PX1, the diffraction resolution needs to be reduced to a relatively low resolution (4.62 Å) to enhance the anomalous signal.

3.3. Substructure determination for sulfur-SAD data

Test case 3 represents a typical case for sulfur-SAD (S-SAD) data. S-SAD is the most common type of native-SAD method. In contrast to SeMet/halogen-SAD, S-SAD utilizes the inherent anomalous signal of the sulfur atoms within the protein for phase determination. However, the anomalous scattering signal of sulfur is relatively weak and the K-absorption edge of sulfur occurs at approximately 5 Å. The quality of diffraction data will decrease significantly at such wavelength. Therefore, in experiments, a wavelength of approximately 2 Å is typically selected for the data collection of S-SAD. In our test case, the wavelengths range from 1.542 Å to 2.755 Å, and the anomalous signal ranges from 5.51 to 9.79 (Table 1). Ultimately, all cases in this set successfully placed $> 70\%$ heavy atoms. Additionally, for the 7O51 case, an extra refinement step (BP3 refinement) was required to separate the single sulfur atom from the disulfide bond (Table 2).

3.4. Substructure determination for RNA/DNA structure

Test case 4 exemplifies the typical challenges encountered in RNA/DNA structure analysis. Given the high flexibility and structural diversity inherent in RNA/DNA molecules, experimental phasing methods often serve as the primary approach for determining their structures. In our test, the default protocol successfully identified 100% of substructure in all cases, clearly demonstrating the robust capabilities of DSAS in handling RNA/DNA structures. Notably, the fact that only 500 strong reflections were required in the 5LQO case underscores the exceptional efficiency and effectiveness of DSAS in this particular context (Table 2).

Table 2. Heavy-atom substructure determination using DSAS.

Test case	PDB entry	Automated parameters setting				Results						[7]Comparison	
		$\mathrm{RES}_{\mathrm{ano}}$	N_{iter}	N_{TF}	w_{best}	[1]nsites	[2]n.c.t.	[3]error	[4]r.m.s.d.	[5]The best CC	[6]Phase error	SHLEXD	*Phenix.hyss*
Test case 1	5COM	2.54	500	1500	0.48	4/4	49	0.22	0.238	34.466	30.14	4/4	3/4
	5XKY	2.3	500	1500	0.28	9/9	29	0.319	0.411	31.067	15.08	9/9	9/9
	8EZS	2.84	500	1300	0.2	7/8	17	0.488	0.546	32.813	31.94	0/8	0/8
	6HP5	3.13	500	1500	0.28	29/30	5	0.338	0.393	39.610	27.14	2/30	4/30
	8PX1	4.62	750	1500	0.48	19/20	364	0.611	0.689	30.173	4978	3/20	2/20
Test case 2	4RYM	2.8	500	1000	0.48	3/3	33	0.607	0.639	31.351	24.99	3/3	3/3
	2RIW	2.35	500	1500	0.38	4/4	22	0.146	0.149	30.264	15.73	0/4	3/4
	5IO8	2.19	500	1500	0.48	11/14	22	0.299	0.336	33.903	28.92	3/18	2/14
Test case 3	8PWN	3.5	500	1300	0.48	16/21	53	0.573	0.606	24.124	39.17	3/21	3/21
	7O51	2.75	500	1300	0.2	13/18[7]	5	0.788	0.872	24.020	21.54	2/18	2/18
	5II7	2.07	500	1500	0.28	15/18	5	0.31	0.357	20.227	29.38	2/18	12/18
Test case 4	5LQO	1.87	500	500	0.48	2/2	60	0.192	0.197	36.637	17.50	1/2	0/2
	3MEI	2.11	500	1000	0.48	2/2	197	0.069	0.081	36.736	13.56	2/2	0/2
	7OW0	1.66	500	1000	0.25	3/3	39	0.158	0.171	36.5445	20.77	0/5	0/3

$\mathrm{RES}_{\mathrm{ano}}$: anomalous resolution;

N_{iter}: the number of iterations for each trial;

N_{TF}: the number of strong reflections;

w_{best} : the optimal percentage of weak reflections.

[1] nsites: the number of sites found in the asymmetric unit (a.u.) compared with published values;

[2] n.c.t.: the number of converging (correct) trials out of 400;

[3] error: the mean error of calculated heavy-atom substructure without BP3 refinement;

[4] r.m.s.d.: the root means square deviation of calculated heavy-atom substructure without BP3 refinement.

[5] The best CC: the highest CC value in the whole iterations.

[6] Phase error: the phase difference between the calculated substructure and the reference substructure.

[7] Comparison: the nsites solved by SHELXD and *Phenix.hyss*.

4. Conclusion and perspectives

As of February 2024, approximately 90% of the macromolecular structures deposited in the PDB have been solved by x-ray crystallography, and SAD phasing remains one of the main structure determination methods. The determination of anomalous scattering substructures is still the main bottleneck of this method. In this work, we developed a new program, DSAS, based on the modified phase-retrieval algorithm. First, the algorithm introduces π-half phase perturbation and the direct-method-based tangent formula into the RAAR algorithm, which effectively improves the success rate and accuracy of the algorithm. Second, thanks to the self-adjusting property of the algorithm parameters, the program can automatically complete the substructure solution without manual intervention, and the entire solution process can be monitored through the GUI. The DSAS has successfully solved some representative SAD experimental data, even those with weak anomalous scattering signals. Therefore, the DSAS is a user-friendly and efficient program for the determination of heavy-atom substructure. In the future, we will continue to optimize the GUI of the DSAS, such as introducing better starting phases from the Patterson function rather than random phases. Furthermore, we will also explore the potential applications of the DSAS program, such as *de novo* small molecule structure determination. Additionally, the command script of DSAS can also function as the core starting module for structure determination pipelines such as *CRANK2*,[29] *IPCAS*,[30] and X^2DF,[31] providing ideal initial phase information for subsequent density modification and model building. It is noteworthy that DSAS will be integrated into a new version of IPCAS in the coming future, aiming to facilitate automatic *de novo* macromolecular structure determination.

Program availability

The code for the DSAS program has been uploaded to GitHub at https://github.com/fuxingke0601/DSAS and Science Data Bank at https://doi.org/10.57760/sciencedb.17996, which are freely available to academic users. The program was written based on the Linux operating platform, which is a commonly used for crystallographic software, making it easy to use with other software. The programs CCP4 should be preinstalled and added to the environment variables of the local system. For more detail, please read the README.md files in the DSAS package.

Acknowledgements

Project supported by the National Natural Science Foundation of China (Grant Nos. 32371280 and T2350011). The electron density maps and structures in Fig. 3 were prepared using PYMOL (https://pymol.org/).

References

[1] Gao X, Shang K, Zhu K, Wang L, Mu Z, Fu X, Yu X, Qin B, Zhu H, Ding W and Cui S 2024 *Nature* **625** 822

[2] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 *Nucleic Acids Res.* **28** 235

[3] Hendrickson W A and Teeter M M 1981 *Nature* **290** 107

[4] Zhang Y, El Omari K, Duman R, Liu S, Haider S, Wagner A, Parkinson G. N and Wei D 2020 *Nucleic Acids Res.* **48** 9886

[5] Schneider B, Sweeney B A, Bateman A, Cerny J, Zok T and Szachniuk M 2023 *Nucleic Acids Res.* **51** 9522

[6] El Omari K, Duman R, Mykhaylyk V, *et al.* 2023 *Commun. Chem.* **6** 219

[7] Zhang S, Wang F, Zhang D, Liu D, Ding W, Springer T A and Song G 2023 *Commun. Biol.* **6** 895

[8] Rose J P and Wang B C 2016 *Arch. Biochem. Biophys.* **602** 80

[9] Weeks C M, DeTitta G T, Miller R and Hauptman H A 1993 *Acta Cryst. D* **49** 179

[10] Schneider T R and Sheldrick G M 2002 *Acta Cryst. D* **58** 1772

[11] Buerger M J 1959 *Vector space: and its application in crystal-structure investigation* (New York: Wiley) pp. 252–266

[12] Bunkóczi G, McCoy A J, Echols N, Grosse-Kunstleve R W, Adams P D, Holton J M, Read R J and Terwilliger T C 2015 *Nat. Methods* **12** 127

[13] Hu M, Gao Z, Zhou Q, Geng Z and Dong Y 2019 *Radiat. Detect. Technol. Methods* **3** 48

[14] Knight S 2000 *Acta Cryst. D* **56** 42

[15] Rius J and Torrelles X 2022 *Acta Cryst. A* **78** 473

[16] Rius J 2011 XLENS_v1: a Computer Program for Solving Crystal Structures from Diffraction Data by Direct Methods *Institut de Ciència de Materials de Barcelona, CSIC*, Spain

[17] Dumas C and van der Lee A 2008 *Acta Cryst. D* **64** 864

[18] Skubák P 2018 *Acta Cryst. D* **74** 117

[19] Luke D R 2005 *Inverse Probl.* **21** 37

[20] Martin A V, Wang F, Loh N D, *et al.* 2012 *Opt. Express* **20** 16650

[21] Oszlányi G and Sütő A 2005 *Acta Cryst. A* **61** 147

[22] Coelho A 2007 *Acta Cryst. A* **63** 400

[23] Sheldrick G 2008 *Acta Cryst. A* **64** 112

[24] Collaborative 1994 *Acta Cryst. D* **50** 760

[25] Oszlányi G and Sütö A 2008 *Acta Cryst. A* **64** 123

[26] Usón I and Sheldrick G M 2018 *Acta Cryst. D* **74** 106

[27] Pannu N S, McCoy A J and Read R. J 2003 *Acta Cryst. D* **59** 1801

[28] Dall'Antonia F and Schneider T R 2006 *J. Appl. Cryst.* **39** 618

[29] Skubák P and Pannu N S 2013 *Nat. Commun.* **4** 2777

[30] Ding W, Zhang T, He Y, Wang J, Wu L, Han P, Zheng C, Gu Y, Zeng L, Hao Q and Fan H 2020 *J. Appl. Cryst.* **53** 253

[31] Ding W, Wang X T and Yi Y Y 2019 *Chin. Phys. B* **28** 116101